A Spatial Modeling Approach for Linguistic Object Data: Analysing dialect sound variations across Great Britain

Shahin Tavakoli,
Statistical Laboratory, University of Cambridge, UK
&
Department of Statistics, University of Warwick, UK,

Davide Pigoli,
Statistical Laboratory, University of Cambridge, UK
&
Department of Mathematics, King's College London, UK,

John A.D. Aston*†
Statistical Laboratory, University of Cambridge, UK,

and

John S. Coleman
Phonetics Laboratory, University of Oxford, UK

May 9, 2018

Abstract

Dialect variation is of considerable interest in linguistics and other social sciences. However, traditionally it has been studied using proxies (transcriptions) rather than

*The authors gratefully acknowledge support from EPSRC grant EP/K021672/2

†Address for correspondence: Professor John Aston, Statistical laboratory, Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, CB3 0WB Cambridge, United Kingdom. Email: j.aston@statslab.cam.ac.uk

acoustic recordings directly. We introduce novel statistical techniques to analyse geolocalised speech recordings and to explore the spatial variation of pronunciations continuously over the region of interest, as opposed to traditional isoglosses, which provide a discrete partition of the region. Data of this type require an explicit modeling of the variation in the mean and the covariance. Usual Euclidean metrics are not appropriate, and we therefore introduce the concept of d-covariance, which allows consistent estimation both in space and at individual locations. We then propose spatial smoothing for these objects which accounts for the possibly non convex geometry of the domain of interest. We apply the proposed method to data from the spoken part of the British National Corpus, deposited at the British Library, London, and we produce maps of the dialect variation over Great Britain. In addition, the methods allow for acoustic reconstruction across the domain of interest, allowing researchers to listen to the statistical analysis.

Keywords: Functional data analysis, acoustic linguistics data, object data analysis, nonparametric smoothing, covariance matrices.

1 Introduction

A better understanding of local dialect variation is of interest both from the point of view of linguistics (how languages evolved in the past, how they became differentiated and how they will develop in the future) and from that of social sciences and demography, the way language is used being both a result of social affiliations and a tool to shape group identification. Dialect variations have long been studied in sociolinguistics by considering textual differences between phonetic transcriptions of the words (see, e.g., Kretszchmar 1996, Nerbonne & Kretzschmar 2003, Nerbonne et al. 2011, and references therein). This focus on written forms reflects a general normative approach towards languages: for cultural and historical reasons, the way we think about them is focused on the written expression of the words, even when thinking of their pronunciations. However, this is more a social artifact than a reality in the population, as there is great variation even within a single region with a claimed "homogeneous" dialect. Indeed, the analysis of speech data highlights that the definition of language is an abstraction that simplifies the reality of speech variability and neglects the continuous geographical spread of spoken varieties, although this does not exclude the presence of some clearly defined boundaries.

In this paper, we develop techniques that explicitly complement this text-based approach; we define methodology not for written or transcription based analysis, but rather by treating the acoustic data directly, by considering sounds as data objects (see Wang et al. 2007, for a definition of data objects). This allows the examination of all forms of variation, including those within groups usually deemed to be homogeneous. To achieve this requires the development of spatially varying statistical models for object data which take into account both the underlying geography, but also the statistical properties of the data (in this case the fact that part of the model requires estimation of quantities which lie on a manifold). This leads to the definition of a new concept of covariance which is statistically consistent over space even under Fréchet type estimation.

We are particularly interested in using information from speech recordings to model the smooth variation of speech characteristics over a geographical region. Since recordings are obtained only in a discrete set of locations, the first step will be the development of a non-parametric smoothing procedure to infer speech characteristics (and plausible speech reconstruction) on the continuous map. Having available replicates from different speakers at each location, we are able to model both the mean and the covariance structure of the speech process at that single location, the latter being highlighted in recent studies (see Aston et al. 2010, Hadjipantelis et al. 2012) as an important feature for language characterization. The model we use to smooth the speech process over the whole geographical region of interest is described in Section 3, with the model based on the concept of using data specific metrics in the analysis.

From a statistical point of view, we develop the concept of spatial object data analysis, and, in particular, the use of d-covariances, that is, covariances that are estimated under a different metric to the usual Euclidean (L^2) one. It has been seen in a variety of applications, particularly diffusion tensor imaging, that even when the use of Euclidean distance is appropriate (and in the case of defining geodesics it may well not be so), it is often sub-optimal in terms of interpretability (see for example Dryden et al. 2009). This is particularly important for the case of spatial smoothing with replicates, as use of the implied Euclidean metric (as is the case for the sample covariance) is not consistent with a spatially smoothed version under another metric, while the Euclidean metric is not valid with general smoothing techniques for positive definite covariances. Thus a new type of covariance will be developed which is statistically consistent. The analysis of the covariance structure is made possible by the presence of replicates of the same sound, uttered by different speakers, in each geographical location. This is an uncommon setting for spatial data analysis which is usually focused on problems where replicates are not available but second-order stationarity can be assumed. The latter is also the setting where most of the recent work on spatial statistics for object data have been developed (see, e.g., Delicado et al. 2010, Gromenko et al. 2012, Menafoglio & Petris 2016). In this work, the need to model the spatial variation of the covariance structure of the speech process led us to choose a non-parametric regression approach to estimate both the mean and the d-covariance of the speech process, in the line of the methods developed for interpolation and smoothing of positive definite matrices (Dryden et al. 2009, Yuan et al. 2012) and for surface smoothing over complex domains (Wood et al. 2008, Sangalli et al. 2013).

We also develop a set of tools to communicate relevant information to linguists. First, we generate colour maps that reflect speech variation in the spirit of isogloss maps (see, e.g., Francis 1959, Upton & Widdowson 2013) but with continuous variation (as opposed to hard boundaries) and using information from speech recordings (as opposed to achieving this via phonetic transcriptions). Moreover, our method allows the resynthesis of a plausible pronunciation for any point in the considered geographical region. We include as supplementary material a few examples of these reconstructed pronunciations for the sound data set described in Section 2.

The paper proceeds as follows. In Section 2, the principles behind using acoustic recordings as the intrinsic data objects, as well as the data set itself, are introduced. Section 3 develops both the concept of d-covariance and the model for spatial data objects based on the d-covariance formulation. Section 4 applies the modeling framework to the British National Corpus data. This data set is a large corpus of acoustic recordings of British English across Great Britain, making it ideal for the comparison of dialects and accents. Finally, Section 5 is a discussion of the work and both its linguistic and statistical relevance. The data, associated R code (R Core Team 2016) to replicate the analysis, supplementary figures, details on the data preprocessing, and technical results concerning the d-covariance and the model are available in the supplementary materials.

2 Sounds As Data Objects

In linguistics, there has recently been a considerable interest in assessing information coming directly from speech recordings (Lehmann 2004, The Functional Phylogenies Group 2012, Pigoli et al. 2014, Hadjipantelis et al. 2015, Coleman et al. 2015) in addition to textual evidence and phonetic transcriptions. While we develop new methodologies that can be applied to a variety of languages and geographical regions, we consider, in particular, the variation of the English language in the United Kingdom. British English is well known to contain a large number of regional dialects, which can have considerable differences between them. Dialect variation is investigated by analysing the spoken part of the British National Corpus (BNC) deposited at the British Library. The digital versions of these recordings are now made available by the Phonetics Laboratory of the University of Oxford (Coleman et al. 2012). These sound data (rather than their phonetic transcriptions) will be directly used to explore British dialects. In particular, for the statistical analysis of speech tokens, it is first necessary to represent sounds in a time-frequency domain and align them in time to account for individual variation in speaking rate. We choose here a Mel Frequency Cepstral Coefficients representation for the speech tokens because of its good performance for speech resynthesis (which will be the final output of our analysis), and because it provides a principled lower dimensional representation of the speech tokens. We now give a more detailed description of the underlying data and their mathematical representation.

2.1 Sound Waves, Spectrograms, and Mel-Frequency Cepstral Coefficients

A one-channel monophonic sound can be represented by a time series $(s(t) : t = 1, \ldots, T)$, where $s(t)$ represents the recording of the air pressure at time t as captured by the microphone. As such, a sound is the variation of air pressure over time. For $t \leq 0$ or $t > T$, we let $s(t) = 0$. We can therefore assume that $s(t)$ is well defined for $t \in \mathbb{Z}$. An example of sound wave is given in Figure 1.

The spectrogram of a sound $(s(t))_{t=1,\ldots,T}$ is a two-dimensional representation $\mathrm{Spec}\,(s)\,(t, \omega)$ of the sound, where $\mathrm{Spec}\,(s)\,(t', \cdot)$ represents the modulus of the discrete Fourier transform of $s(t)$ in a neighborhood of t'. Mathematically, if $W(x), x \in \mathbb{R}$ in a window function with support $[-1, 1]$, then for any positive integer $M, w_M(t) = W(2t/M), t \in \mathbb{Z}$ is a window of width M, and

$$\mathrm{Spec}\,(s)\,(t, \omega) = \left| \sum_{u=1}^{T} s(t-u) w_M(u) \exp(-i\omega u) \right|, \quad t = 1, \ldots, T, \omega \in [0, 2\pi].$$

The function $u \mapsto s(t-u)w_M(u)$ is a windowed version of s around t. For computational efficiency, the spectrogram is computed at the Fourier frequencies $\omega \in \{2\pi k/N\}$, where $N \geq T$ is highly composite (usually a power of 2), using the fast Fourier transform (FFT; Cooley & Tukey 1965). The window width M is typically chosen to correspond to a segment of length ranging from 5 to 20 milliseconds ($M = 80$ or 320 at 16Khz). From now on, we shall call the spectrogram of s the $T \times N$ matrix with entries $\mathrm{Spec}\,(s)\,(t, \omega_k), t = 1, \ldots, T, \omega_k = 2\pi k/N, k = 0, \ldots, N-1$.

A low dimensional time-frequency representation of the sound wave, often used in speech recognition and speech synthesis, are the *Mel-frequency cepstral coefficients*, or *MFCC*. The computation of the MFCC is done in two steps. First, the *Mel spectrogram*, a filtered version of the spectrogram is computed,

$$\mathrm{MelSpec}\,(s)\,(t,f) = \sum_{k=0}^{N-1} \mathrm{Spec}\,(s)\,(t, 2\pi k/N) b_{f,k}, \quad f = 0, \ldots, F,$$

where $(b_{f,k})_{k=0,\ldots,N-1}, f = 0, \ldots, F$ is the so-called Mel-scale filter bank (an example of a Mel-scale filter bank is given in Gold et al. 2011) with F filters, which is believed to mimic the human ear auditory system. Then, the MFCC corresponds to the first $M \leq F$ coefficients of the inverse Fourier transform of the Mel spectrogram:

$$\mathrm{MFCC}_s(t,m) = \frac{1}{F} \sum_{f=0}^{F} \log\left(\mathrm{MelSpec}\,(s)\,(t,f)\right) \exp\left[i(2\pi(m-1)/(F+1))f\right], \quad m = 1, \ldots, M.$$

An additional reason to prefer MFCC over spectrograms is that each coefficient is associated to a frequency band, and therefore the MFCCs are more robust to small misalignments in frequency when comparing multiple speakers or sounds. Since the MFCCs are assumed to be smooth in t, we shall from now on assume that $t \in [0, 1]$, where it is implicitly assumed that the integer t are replaced by t/T and interpolated. An example of MFCC is given in Figure 1.

Note that there exist many modifications and variations of this definition of MFCC in the literature, as authors seek improvements in the performance of implemented speech recognition or parametric speech synthesis systems. Since one of the goals of this paper is the resynthesis of sounds after inference, we shall use the definition and computational implementation of the MFCC proposed in Erro et al. (2011, 2014) as it yields high-quality, natural sounding resynthesised speech. However, the underlying principles are the same. For simplicity, we will refer in the following to this modified version as MFCC.

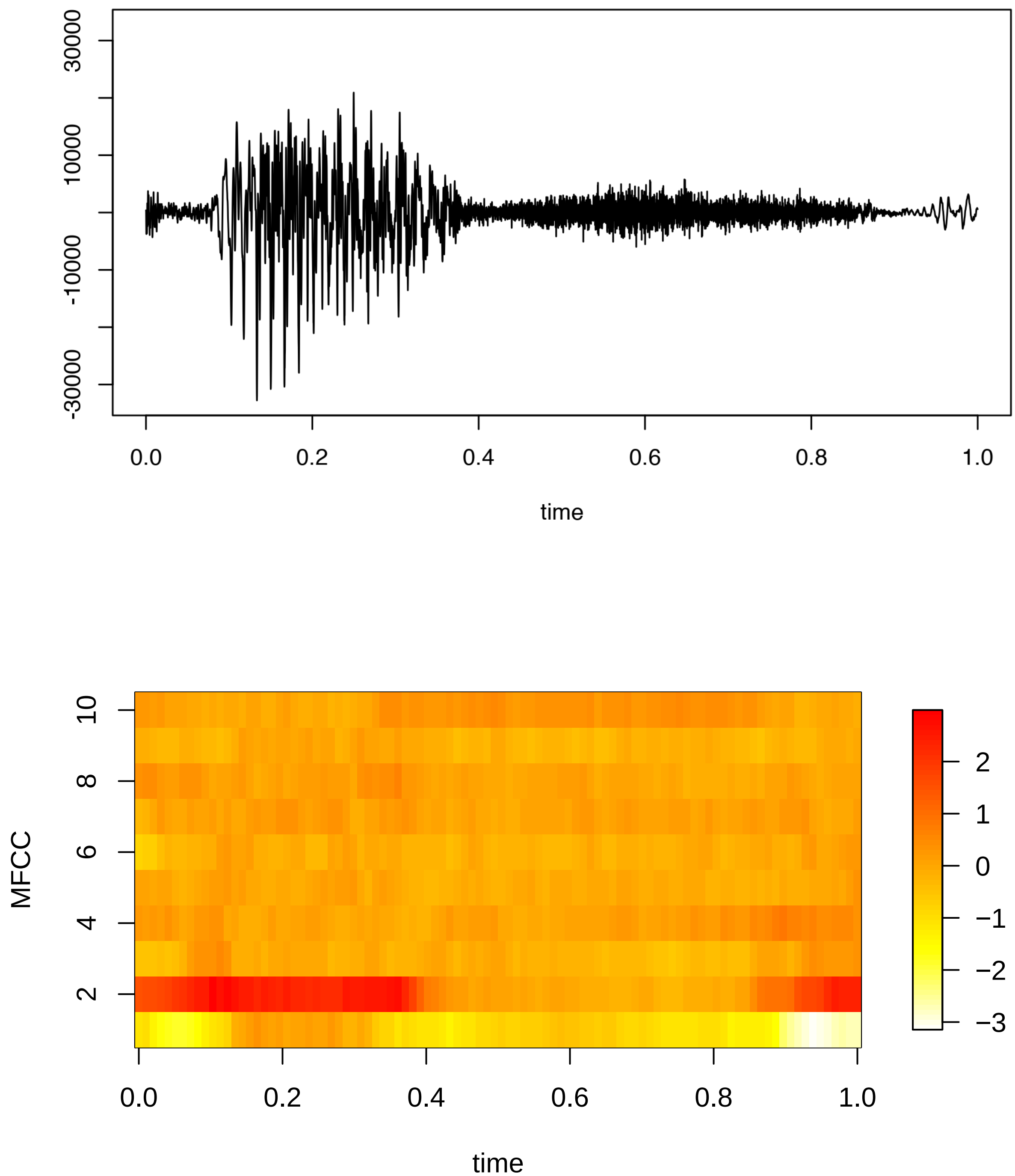


Figure 1: Top figure: sound wave $s(t)$ of the word 'last'. Bottom figure: corresponding $\mathrm{MFCC}_s(t, m)$. Both time scales have been normalized to $[0, 1]$.

2.2 The British National Corpus

The raw data consist of the audio British National Corpus (BNC) recordings (131 GB of data, 16 Bit 16 kHz one-channel .wav files, roughly 1100 hours of recording, publicly available at `http://www.phon.ox.ac.uk/AudioBNC`). These are mainly recordings of natural speech in typically noisy environments, with low recording amplitude (signal-to-noise ratio). Segmentation information about the words pronounced in the audio files were also provided (in TextGrid format), with the XML edition of the BNC (4.4 GB of files) containing transcriptions of the words spoken in the audio BNC recordings, along with contextual information (anonymized speaker identification, information about the speakers, location of the recording).

For the purposes of the current paper, we restricted ourselves to the analysis of sounds of the vowel "a" present in the following list of words:

`class, glass, grass, past, last, brass, blast, ask, cast, fast, pass`. (2.1)

The vowels in these words are pronounced in the same (geographically consistent) way and therefore we can consider them as the replicates of the same sound. We denote this as the "class" dataset. In Great Britain, this vowel is considered prototypical of the distinction between northern and southern accents: in the Midlands, North and South-West these words have a short, open front vowel [a] as in "pat", whereas in the South and South-East they have a long back vowel [ɑ] ("aah"), similar to the vowel in "part". The purpose of our work is the spatial analysis of sounds, and as such we needed to assign to each recorded sounds to the geographical location of the speaker's origin. We therefore removed sounds of speakers with missing or vague location information, and sounds corresponding to speakers who where trained to speak in a specified fashion (such as TV or radio presenters). For each speaker, we then used the corresponding recording location (the variable `placenamecleaned`) as a surrogate for the speaker's origin, provided this was unique. If there were multiple recording locations for a given speaker, the location corresponding to the `locale` variable "home" or "at home" was taken as the location of origin. If no such location existed, the sounds corresponding to the speaker were discarded. It should be noted that although we assume the speaker's accent to be representative of his recording location, there is unfortunately no data about the origin of the speakers to corroborate this assumption.

After this process, we obtained 4816 sound tokens from 110 distinct geographical locations in Great Britain, with 1993 distinct speaker-vowel (from distinct word) combinations. About 46% of the speaker-vowel combinations appear more than once, 61 speaker-vowel combinations appear at least 10 times, and there is a speaker-vowel combination that is repeated 34 times in the dataset. While the model we will use in Section 3 could be extended to a random effects type setup, to account for speaker repetition, we prefer to concentrate on the simpler model to aid understanding, particularly as most speaker vowel combinations only appear once.

These vowel sounds were then transformed into MFCCs, with $M = 10$. For each word w in our list of words (2.1), we aligned the MFCCs of sounds corresponding to w by registering their first coefficient (which corresponds to relative volume) using the Fisher-Rao metric (R package `fdasrvf`; see Srivastava et al. 2011, Tucker et al. 2013, Wu & Srivastava 2014) and then extracted the segment associated to the vowel for each word, and linearly rescaled its time to the unit interval $[0, 1]$. Furthermore, we centered the first cepstral coefficients of all the data to remove differences in recording volume. For further information concerning the preprocessing and alignment, see the Online Supplement.

We note that the duration of the vowel sounds is lost in the preprocessing step, and as such, we present a spatial map of the vowel durations in the Online Supplement (Figure S5). Indeed, the vowel duration is a one scalar summary of the sound of the vowel, which although useful does not capture considerable additional qualitative information contained in the vowel sound MFCCs, even after time alignment.

3 Model and Estimation

As mentioned earlier, previous works have identified the covariance structure between frequencies as an important feature of the speech process that characterises languages (Aston et al. 2010, Hadjipantelis et al. 2012, Pigoli et al. 2014, Hadjipantelis et al. 2015). We therefore have good reasons to expect the covariance between MFCCs—which are related to the energy in each frequency band—to be associated with dialect characteristics, and we want to allow for it to vary geographically. The investigation of the best metric for interpolation or extrapolation of covariance matrices or operators has recently generated much work (Arsigny et al. 2007, Dryden et al. 2009, Yuan et al. 2012, Carmichael et al. 2013). The use of a metric different from the Euclidean metric in the analysis leads to the formulation of a more general concept of co-variability: the d-covariance. In the following, we explain why we need to introduce this new concept and the role it plays in the definition of the model for speech variation presented in Section 3.2.

3.1 d-covariances

Interpolation of covariance matrices under the usual Euclidean metric, although yielding valid covariances, suffers from artifacts, such as swelling (e.g. Arsigny et al. 2006). Extrapolation of covariances under the Euclidean metric, on the other hand, is not even guaranteed to give valid covariances. For this reason, several other metrics on the spaces of symmetric positive semi-definite matrices have been studied, and have been shown to be useful for interpolation or extrapolation of covariances. For instance, the Euclidean average $\overline{C} = n^{-1} \sum_{i=1}^{n} C_i$ of covariance matrices $C_1, \ldots, C_n$ can be reformulated as the solution to the variational problem

$$\min_{\Omega} \sum_{i=1}^{n} d_E^2(\Omega, C_i),$$

where d_E denotes the Euclidean distance. In other words, $\overline{C}$ is the Fréchet mean of $C_1, \ldots, C_n$ under d_E. Therefore the average of covariances $C_1, \ldots, C_n$ under another metric d can be defined as their Fréchet mean under d.

While covariance interpolation or extrapolation under various metrics is useful (for example in the case of spatial smoothing; see e.g. Yuan et al. 2012), it is only valid when treating the covariances C_i as the observation units. However, the covariances C_i are estimators of unknown true covariances, and have therefore an intrinsic estimation error. Since the true covariance of a random vector $X \in \mathbb{R}^p$ can be defined as the solution of the variational problem

$$\min_{\Omega} \mathbb{E}\, d_E^2 \left(\Omega, (X-\mu)(X-\mu)^\top\right),$$

the sample covariance can be viewed as an analogue sample-based variational problem *based on the Euclidean metric* d_E. When using a metric different than d_E for spatial smoothing of sample covariances, a consistency problem arises due to the two different metrics used in the variational problem and the smoothing problem, and the resulting estimator is biased. A one-dimensional example illustrating this is given in Section S6 of the Online Supplement.

For this reason, we introduce the concept of d-covariance that stems from recent developments on the inference for covariance operators (see Arsigny et al. 2006, Dryden et al. 2009, Kraus & Panaretos 2012, Pigoli et al. 2014, Petersen & Müller 2016) , where d is a metric on the space of $p \times p$ symmetric positive semi-definite matrices $\mathcal{S}_p$ that is used for the spatial smoothing. The d-covariance of a random vector $X \in \mathbb{R}^p$ is denoted $\operatorname{cov}_d(X)$, and defined by

$$\operatorname{cov}_d(X) = \underset{\Omega \in \mathcal{S}_p}{\operatorname{argmin}}\, \mathbb{E}\, d^2\left((X-\mu)(X-\mu)^\top, \Omega\right),$$

where $\mu = \mathbb{E}\, X$, and provided the right-hand side is well defined.

In this paper, we shall use the square-root metric d_S on the space of symmetric positive semi-definite matrices, defined by $d_S(B,C) = \left|\!\left|\!\left| \sqrt{B} - \sqrt{C} \right|\!\right|\!\right|$, where $\sqrt{B}$, also written $B^{1/2}$, is the unique square root of B (meaning that it is the unique matrix D that satisfies $DD = B$; see Section S1 of the Online Supplement), and $|||\cdot|||$ is the Frobenius norm. As can be seen in the one-dimensional example given in Section S6 of the Online Supplement, using the same metric d_S for both the definition of the co-variation and the spatial smoothing yields an estimator that is less biased than the one obtained by spatial smoothing of the usual (Euclidean) covariance with d_S. Let $|\cdot|$ denote the Euclidean norm on $\mathbb{R}^p$, i.e. $|x| = \sqrt{x^\mathsf{T} x}$. The following Proposition gives an explicit formula for the d_S-covariance.

Proposition 3.1. *Let $X \in \mathbb{R}^p$ be random element with $\mathbb{E}\,|X| < \infty$ and mean $\mu = \mathbb{E}\,X$. Then* $\operatorname{cov}_{d_S}(X) = \mathbb{E}\left[\sqrt{(X-\mu)(X-\mu)^\mathsf{T}}\right]^2$.

Notice in particular that we do not need second moments for the d_S-covariance to exist, which is due to the fact that $\left|\!\left|\!\left|\sqrt{XX^\mathsf{T}}\right|\!\right|\!\right| = |X|$. Since there is an explicit formula for the square-root of symmetric positive semi-definite matrices of rank one, namely $\sqrt{xx^\mathsf{T}} = xx^\mathsf{T}/|x|$, for $x \in \mathbb{R}^p, x \neq 0$, we can rewrite the d_S-covariance of X as

$$\operatorname{cov}_{d_S}(X) = \mathbb{E}\left[\frac{(X-\mu)(X-\mu)^\mathsf{T}}{|X-\mu|}\right]^2, \tag{3.1}$$

where the expression inside the expectation is understood to be equal to zero if $X = \mu$. The denominator in (3.1) reveals that the square-root of the d_S-covariance can be viewed as a regularized version of the usual covariance. Furthermore, it also reveals that unlike the Euclidean covariance, the d_S-covariance does not behave in the usual way under linear transformations: $\operatorname{cov}_{d_S}(AX) \neq A\operatorname{cov}_{d_S}(X)A^\mathsf{T}$ for general linear transformations A. However, if we introduce the new families of square-root semi-metrics $d_{S,A}(C,D) = \left|\!\left|\!\left|\sqrt{ACA^\mathsf{T}} - \sqrt{ADA^\mathsf{T}}\right|\!\right|\!\right|$, where C, D are $p \times p$ symmetric positive semi-definite matrices and A is a $n \times p$ matrix, we have the following result, proved in Section S2 of the Online Supplement.

Proposition 3.2. *Let A be a $n \times p$ matrix, and $X \in \mathbb{R}^p$ be a random element with $\mathbb{E}\,|X| < \infty$. Then $\operatorname{cov}_{d_S}(AX) = A\operatorname{cov}_{d_{S,A}}(X)A^\mathsf{T}$. In the special case where $A^\mathsf{T}A = I$, the identity matrix, we have $\operatorname{cov}_{d_S}(AX) = A\operatorname{cov}_{d_S}(X)A^\mathsf{T}$.*

This means that the d_S-covariance of a linear transformation of X is given by a transformation of the d-covariance of X under a metric related to the linear transformation. In particular, the entries of a d_S-covariance do not correspond to the d_S-covariance of corresponding entries of the random vector. This is analogous to partial correlation. Furthermore, Proposition 3.2 tells us that the d_S-covariance is rotation equivariant. Note also that the d_S-covariance is a measure of spread, and that other measure of spreads have been proposed for multivariate data or functional data (Locantore et al. 1999, Gervini 2008, Kraus & Panaretos 2012), motivated from a robustness perspective.

3.2 A model for spatially varying speech object data

We are now ready to define the model for speech variation for the analysis of dialect data. We wish to have a model which can spatially vary both in terms of a mean function but also in terms of covariance, as we will have replicates at individual spatial locations. We therefore assume the following model:

$$Y_{lj}(t) = m(X_l, t) + \varepsilon_{lj}(t), \quad l = 1, \dots, L; j = 1, \dots, n_l, \tag{3.2}$$

where $Y_{lj}(t) \in \mathbb{R}^p$ is the vector of the first p Mel-frequency cepstral coefficient (MFCC) at time $t \in [0, 1]$ of the recording lj, X_l corresponds to the spatial location of the observations $Y_{lj}, j = 1, \dots, n_l$, recorded in latitude/longitude coordinates, i.e. $X_l \in \mathcal{E}$, where $\mathcal{E} \subset (-90, 90] \times (-180, 180]$ is the spatial domain, and will denote Great Britain in the application of Section 4. The *spatial MFCC* is the function $x \mapsto m(x, \cdot) \in L^2([0, 1], \mathbb{R}^p)$, mapping a spatial location $x \in \mathcal{E}$ to its corresponding mean MFCC.

The term $\varepsilon_{lj} \in L^2([0, 1], \mathbb{R}^p)$ is an error term. We assume that for each $l = 1, \dots, L$, $\varepsilon_{lj} \overset{\text{iid}}{\sim} \varepsilon(X_l), j = 1, \dots, n_l$, and that the $\varepsilon_{lj}s$ are all independent. Indeed, this is a valid assumption since we have replicates for each location X_l, and a scatterplot of the pairwise distances between the errors against their geographical distances does not reveal any spatial dependence (see Figure S6 of the Online Supplement). The process $\varepsilon(\cdot) : \mathcal{E} \to L^2([0, 1], \mathbb{R}^p)$ is assumed to have mean zero, $\mathbb{E}\,\varepsilon = 0$, and we denote its d_S-covariance by $\Omega(x, t) = \mathrm{cov}_{d_S}(\varepsilon(x, t))$, where we write $\varepsilon(x, t)$ for $\varepsilon(x)(t)$. This implies in particular that $\mathrm{cov}_{d_S}(Y_{lj}(t)) = \Omega(X_l, t)$. While traditionally $\Omega(X_l, t)$ would be defined as the covariance matrix of Y_{lj}, by assuming that $\mathbb{E}\,\varepsilon_{lj}(t) = 0$ for all t and $\mathbb{E}\left[\varepsilon(t)\varepsilon(t)^\top\right]$ is the identity, we define here $\Omega(X_l, t)$ to be the d_S-covariance of Y_{lj}, where d_S is the square-root metric, because we shall be smoothing spatially using the metric d_S. Recalling that $\mathcal{S}_p \subset \mathbb{R}^{p\times p}$ is the space of symmetric positive semi-definite $p \times p$ real matrices, the function $x \mapsto \Omega(x, \cdot) \in L^2([0, 1], \mathcal{S}_p)$, maps a spatial location $x \in \mathcal{E}$ to a time-varying symmetric positive semi-definite matrix at that location.

Given the observations $\{Y_{lj}(t), X_l\}$, we want to estimate a smooth field $\widehat{m}(x, t)$ for the mean of the speech process, and a smooth field $\widehat{\Omega}(x, t)$ for the (time-dependent) d_S-covariance between MFCCs coefficients.

3.3 Estimation of the mean MFCCs field

In this section, we will be dealing with the estimation of the mean MFCC field m, and therefore the natural metric in this case to consider is the Euclidean (L^2) distance. However, when we consider the geographical distance, the natural metric is the geodesic distance, which we will approximate by graph distance $\mathrm{d_g}(\cdot, \cdot)$ on a constructed triangular mesh over the region of interest.

We propose to fit the mean MFCC field using a local constant estimator which minimizes a weighted mean square fit criterion. Let $K : \mathbb{R} \to [0, \infty)$ denote a continuous and bounded density function, and let $K_h(s) = K(s/h)/h^2$. At the location x, the estimate of the mean MFCC is $\hat{m}(x) \in L^2([0, 1], \mathbb{R}^p)$ which minimizes

$$\sum_{l=1}^{L} \sum_{j=1}^{n_l} K_h\left(\mathrm{d_g}(x, X_l)\right) \frac{\|Y_{lj} - \hat{m}(x)\|^2}{\hat{\sigma}^2(X_l)}, \tag{3.3}$$

where $\|\cdot\|$ is the usual norm in $L^2([0, 1], \mathbb{R}^p)$, i.e. $\|f\|^2 = \int_0^1 |f(t)|^2 dt$ for $f \in L^2([0, 1], \mathbb{R}^p)$, and $\mathrm{d_g}(x, X_l)$ is the distance on the map between x and X_l. The denominator is a normalizing factor that compensates for possible heteroscedasticity in the MFCC field using the total variability of the residuals, $\hat{\sigma}^2(X_l) = n_l^{-1} \sum_{j=1}^{n_l} \left\|Y_{lj} - \overline{Y}_l\right\|^2$, where $\overline{Y}_l = n_l^{-1} \sum_{j=1}^{n_l} Y_{lj}$. The minimizer of the fit criterion (3.3) is a Nadaraya–Watson type estimator, given by convex combination of the average MFCCs at each location, i.e.

$$\hat{m}(x) = \sum_{l=1}^{L} w_l(x) \overline{Y}_l, \tag{3.4}$$

where

$$w_l(x) = \tilde{w}_l(x) / \sum_{l'=1}^{L} \tilde{w}_{l'}(x) \quad \& \quad \tilde{w}_l(x) = n_l \cdot K_h \left(\mathrm{d_g}(x, X_l) \right) / \hat{\sigma}^2(X_l)$$

Possible strategies for the choice of the bandwidth h are discussed in Section 4.1. It may be argued that using a higher order local polynomial estimator in place of (3.3) can reduce the bias of the estimator, and there exists methods to perform local linear smoothing when only pairwise distances between the covariates are available (Baíllo & Grané 2009, Boj et al. 2010, 2016). We leave this extension as a future avenue of research.

3.4 d_S-Covariance Field Estimation

In this section, we extend the kernel smoother to estimate the smooth d_S-covariance field Ω. The natural metric to be used for the smoothing in this case is the square root metric d_S as this indeed avoids inconsistencies between estimation of the d_S-covariance in the observed locations and estimation of the spatially smooth field, as we show in Section 3.5. Moreover, the square root metric is well-defined for singular matrices, a property that will be needed for the application to the BNC data in Section 4. Indeed, locations with small number of observations are expected to have d_S-covariance between MFCCs that are not full rank. We also propose to use a locally constant estimator of the covariance field to allow for the non-convex domain, as discussed in the previous section.

At the point x, the estimated covariance $\hat{\Omega}(x, \cdot) \in L^2\left([0,1], \mathcal{S}_p\right)$ is the minimizer of the following fit criterion:

$$\sum_{l=1}^{L} K_h \left(\mathrm{d_g}(x, X_l) \right) \int_0^1 d_S^2(\breve{\Omega}_l(t), \hat{\Omega}(x,t)) dt, \tag{3.5}$$

where h is a smoothing parameter, and $\breve{\Omega}_l \in L^2([0,1], \mathcal{S}_p)$ is the sample d_S-covariance at location X_l, defined as

$$\breve{\Omega}_l(t) = \underset{\Omega \in \mathcal{S}_p}{\operatorname{argmin}} \ \frac{1}{n_l} \sum_{i=1}^{n_l} d_S^2 \left(\Omega, (Y_{li}(t) - \overline{Y}_l(t))(Y_{li}(t) - \overline{Y}_l(t))^\mathsf{T} \right), \quad \text{for each } t \in [0,1].$$
$$= \left[\frac{1}{n_l} \sum_{i=1}^{n_l} \sqrt{(Y_{li}(t) - \overline{Y}_l(t))(Y_{li}(t) - \overline{Y}_l(t))^\mathsf{T}} \right]^2$$

It is not difficult to show (see Lemma S2.1 of the Online Supplement) that the minimizer of (3.5) is given by

$$\hat{\Omega}(x,t) = \left[\sum_{l=1}^{L} w_l(x) \sqrt{\breve{\Omega}_l(t)} \right]^2, \tag{3.6}$$

where

$$w_l(x) = \tilde{w}_l(x) / \sum_{l'=1}^{L} \tilde{w}_{l'}(x) \quad \& \quad \tilde{w}_l(x) = K_h\left(\mathrm{d_g}(x, X_l)\right). \tag{3.7}$$

Equation (3.6) reveals that $\hat{\Omega}(x)$ is the square of a Nadaraya–Watson estimator in the square-root space.

3.5 Consistency of Smoothing with the Square Root Distance

In this section, we study the properties of the estimator for the d_S-covariance smooth field $\Omega(x,t)$. This is a non-standard smoothing problem, which poses a few theoretical challenges due to the non-Euclidean metric involved, and to the fact that we want to control the estimation error uniformly in the time index. Moreover, this gives us the opportunity to show how it is possible to account for the use of the geographical distance $\mathrm{d_g}$ in the kernel smoothing. The estimator for the mean field $m(x,t)$ uses the Euclidean (L^2) metric and its properties can therefore be studied using similar arguments, in particular using the results in Section S2 of the Online Supplement.

This first result, proved in Section S2 of the Online Supplement, shows that under mild assumptions, the sample d_S-covariance is a $\sqrt{n}$-consistent estimator of the d_S-covariance.

Proposition 3.3. *Let* $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} Y \in \mathbb{R}^p$ *be random vectors with* $\mu = \mathbb{E}\, Y$ *and* $\mathbb{E}\, |Y|^2 < \infty$. *Let* $\overline{Y} = (Y_1 + \cdots + Y_n)/n$, *and*

$$\breve{\Omega} = \left(\frac{1}{n} \sum_{i=1}^{n} \sqrt{(Y_i - \overline{Y})(Y_i - \overline{Y})^\intercal} \right)^2 ,$$

this being the explicit expression for the sample d_S*-covariance. Then* $d_S(\breve{\Omega}, \mathrm{cov}_{d_S}(Y)) \leq \kappa_p \sqrt{\mathbb{E}\, |Y - \mu|^2 / n}$, *where* κ_p *is a constant depending only on the dimension.*

In particular, $\breve{\Omega} = \mathrm{cov}_{d_S}(Y) + O_{\mathbb{P}}(n^{-1/2})$. We now introduce some conditions used in proving the consistency of the smooth d_S-covariance field.

Condition 3.4.

(1) The kernel $K : \mathbb{R} \to [0, \infty)$ *is a continuous probability density, with* $\int_0^\infty s^3 K(s) ds < \infty$. *Assume also that* K *is decreasing, i.e.* $0 \leq s \leq t \implies K(s) \geq K(t)$.

(2) There exists constants $0 < c_1 < c_2$ *such that* $c_1 |x - y| \leq \mathrm{d_g}(x, y) \leq c_2 |x - y|$.

Condition 3.4 (1) is a standard condition on the kernel function, which is in particular satisfied by the Gaussian kernel we use in Section 4. Condition 3.4 (2) states that the graph distance is (metric) equivalent to the Euclidean distance on $\mathcal{E}$. The following condition on the sampling density is standard.

Condition 3.5. *The density of the observation locations* $X_1, \ldots, X_L \in \mathcal{E}$, $f : \mathcal{E} \to \mathbb{R}$ *is continuous, and* $\sup_{x \in \mathcal{E}} f(x) < \infty$.

Recall that $\Omega(x, t) = \mathrm{cov}_{d_S}(\varepsilon(x, t))$. We are going to assume the following regularity conditions on the error process ε.

Condition 3.6.

(1) $n_l \geq c_0 n$ *for all* $l = 1, \ldots, L$ *for some* $c_0 > 0$.

(2) $\sqrt{\Omega(\cdot,t)} : \mathcal{E} \to \mathcal{S}_p$ *is* C^1 *(with respect to the Hilbert–Schmidt norm), and* $\sup_{x\in\mathcal{E},t\in[0,1]} \left| \frac{\partial\left[\sqrt{\Omega(x,t)}\right]_{rs}}{\partial x}(x) \right| < \infty$, *where* $[A]_{rs}$ *is the* rs*-th entry of the matrix* A.

(3) $\sup_{x\in\mathcal{E},t\in[0,1]} \mathbb{E}\left|\varepsilon(x,t)\right|^2 < \infty$.

Condition 3.6 (1) states that asymptotically, the number of observations per locations is of the same order. Condition 3.6 (2) is a (pointwise in time) smoothness condition on the d_S-covariance field. Condition 3.6 (3) assumes that the second moment of the error field is uniformly bounded. The second moment is needed to establish the rate of convergence, whereas the uniform bound is for the control of the smoothing error uniformly in time.

We can now state the result on the consistency of the smoothed d_S-covariance field, whose proof is in Section S2 of the Online Supplement.

Theorem 3.7. *Assume model* (3.2) *with conditions 3.4, 3.5 and 3.6 holds, and* $L \to \infty, h \to 0, Lh \to \infty, n \to \infty$. *Then, for any* $x \in \mathcal{E}$ *in the interior of* $\mathcal{E}$ *such that* $f(x) > 0$, *we have*

$$\mathbb{E}_X\, d_S(\hat{\Omega}(x,t), \Omega(x,t)) = O_{\mathbb{P}}(n^{-1/2}) + O_{\mathbb{P}}\left(\sqrt{h^2 + \frac{1}{nLh^2}}\right),$$

where the stochastic term is uniform in t, *and* $\mathbb{E}_X$ *is the expectation conditional on* $X_1, \ldots, X_L$.

The first error term comes from the fact that we are using the sample mean in place of the true mean in the computation of the sample d_S-covariance, while the second error term is a bias plus variance decomposition. Notice that the n in the variance term is unusual, and is related to the estimation error of d_S-covariances at the observation locations. In particular, the variance is inversely proportional to the number of observations per location, regardless of L and h.

4 Analysis of sound data from the BNC

We apply here the proposed method to the "class" dataset described in Section 2. As mentioned, the sound tokens come from 110 distinct locations within Great Britain, which are indicated on the geographical map in Figure 2. Also shown is the triangulation used for the smoothing, where the internal nodes contain the locations of the observations. It can be seen that the observed locations are irregularly spaced in the region, with high density of the observations around London and other large cities and very sparse observations in Wales and central Southern England, for example. In particular, only three locations are available in Scotland and therefore we will not draw strong conclusions about the dialect variation in that country. Figures S3 and S4 in the Online Supplement show the counties and regions of Great Britain.

While the method allows for a smooth reconstruction of the sound from the mean MFCC (and a few of these reconstructed sounds for the vowel described above can be found as Supplementary Material), we want also to represent the sound variations (and those of the their d_S-covariance) on a map to be able to explore dialect variations. We need therefore to reduce the dimensionality of the data object. Among the possible alternatives, we choose to project the mean smooth field onto the principal components obtained from the original data because this allows the comparison of the projections of the smooth field estimated using different choices of bandwidth parameters. Concerning the d_S-covariance smoothed field, its variation will be explored by considering the pairwise distances of the estimated d_S-covariance field, and comparison of these distances with those obtained under the assumption that there is no spatial variation in the d_S-covariance field. We will use a Gaussian kernel for all the results of this Section, and we will now discuss the choice of the smoothing parameters.

4.1 Choice of Smoothing Parameters

4.1.1 Varying Bandwidths

In both the estimator for the mean and the covariance, a bandwidth varying with the geographical location can be used. This is particularly important when the locations of the observations are irregularly spaced in the region of interest, as is the case for the "class"

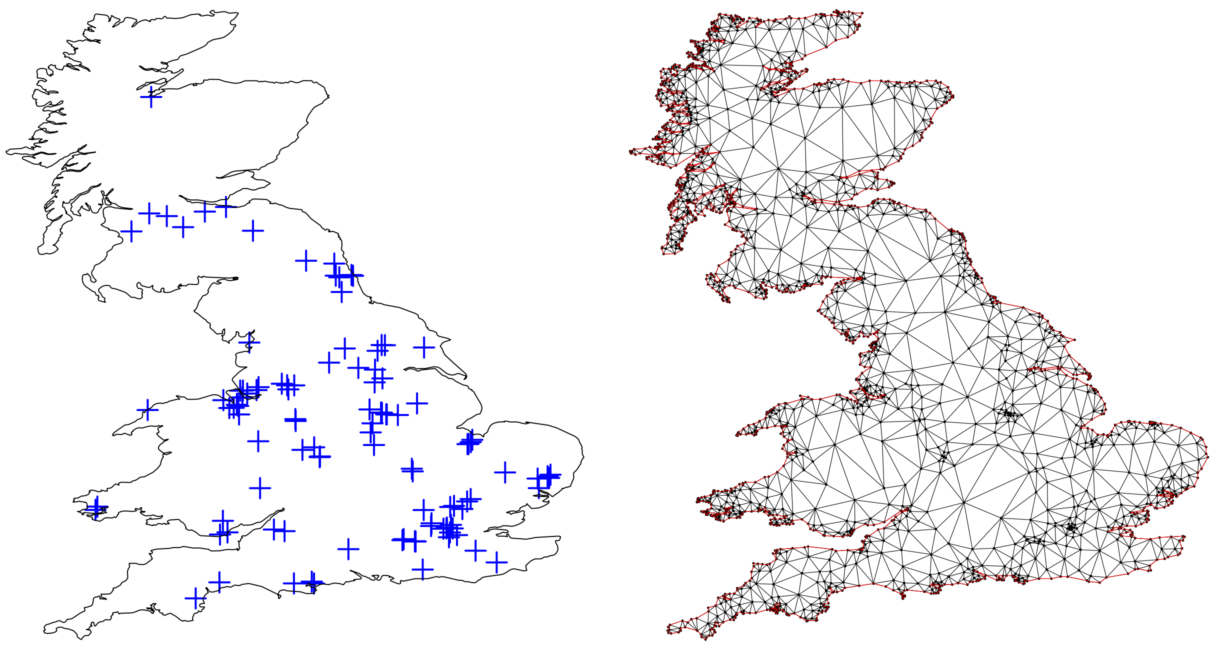


Figure 2: Geographical locations where data are available in the “class” dataset (left) and triangulation of Great Britain (right).

dataset, where the use of a constant bandwidth would lead to over- or undersmoothed estimates.

A possible approach is to adapt the bandwidth to the density of the observations using the distance from the k-th nearest *location* to modulate the global bandwidth, i.e.

$$\hbar(x) = h \cdot \Delta_L(x, k), \tag{4.1}$$

where $\Delta_L(x, k)$ is the geographical distance between x and the k-th nearest location to x. This *k-th-nearest locations varying bandwidth* adjusts the bandwidth to the density of the observed locations, thus guaranteeing that information from comparable numbers of observed locations are used in the estimation at each point. However, if there is a large variability between number of observations at different locations, one may prefer to adjust to the number of observations. We can then define a *k-th-nearest observations varying bandwidth* as

$$\hbar(x) = h \cdot \tilde{\Delta}(x, k), \tag{4.2}$$

where $\tilde{\Delta}(x, k)$ is the least distance from x within which there are at least k observations, i.e.

$$\sum_{l\,:\,d(x,X_l)<\tilde{\Delta}(x,k)} n_l < k \qquad \text{and} \qquad \sum_{l\,:\,d(x,X_l)\leq\tilde{\Delta}(x,k)} n_l \geq k.$$

A third alternative would be to simply using a fixed bandwidth $\hbar(x) = h$ for all x, but this leads to the problem of oversmoothing in the regions with denser observations, as mentioned above.

The idea of adjusting the bandwidth on the basis of observation density is well known in non-parametric regression (see e.g. Fan & Gijbels 1995), but the difficulty in estimating the bivariate density with a relatively small numbers of observations led us to prefer the use of the distance from the k-th nearest neighbour as proxy for the inverse of the density of the observations, this distance being expected to be small in high density regions and large in low density regions.

The expressions of the bandwidth in (4.1) and (4.2) contain two parameters that need to be chosen: the number k of nearest neighbours to be used to adapt the bandwidth and the global smoothing parameter h. These can be chosen by cross-validation, as described in the next Section.

4.1.2 Cross-validation for varying bandwidth parameters

The choice of the parameters k and h for the varying bandwidths (4.1) and (4.2) can be guided by estimating the prediction error as a function of such parameters using a cross-validation procedure. We propose here to use a leave-one-location-out cross validation for the choice of the parameters k and h. For the mean field, the cross-validation is defined by

$$\text{mean.cv}(k,h) = \sum_{l=1}^{L} \frac{\left\|\overline{Y}_l - \hat{m}_{-l}(X_l)\right\|^2}{\hat{\sigma}^2(X_l)},$$

where $\hat{m}_{-l}$ is the estimate of the MFCC field obtained without all the MFCCs observed at location X_l. Analogously, we can define a cross-validation error for the d_S-covariance estimator as

$$\text{cov.cv}(k,h) = \sum_{l=1}^{L} \int_0^1 d_S^2(\widehat{\Omega}_{-l}(X_l)(t), \breve{\Omega}_l(t))\mathrm{d}t,$$

where $\widehat{\Omega}_{-l}(X_l)$ is the prediction for the d_S-covariance at location X_l obtained from (3.6) without the observations at location X_l, and $\breve{\Omega}_l$ is the sample d_S-covariance at X_l. It is however important also to explore the results visually, using the strategies described in Section 4.2, for different values of the smoothing parameters to be sure that the chosen parameters are not leading to oversmoothing or overfitting.

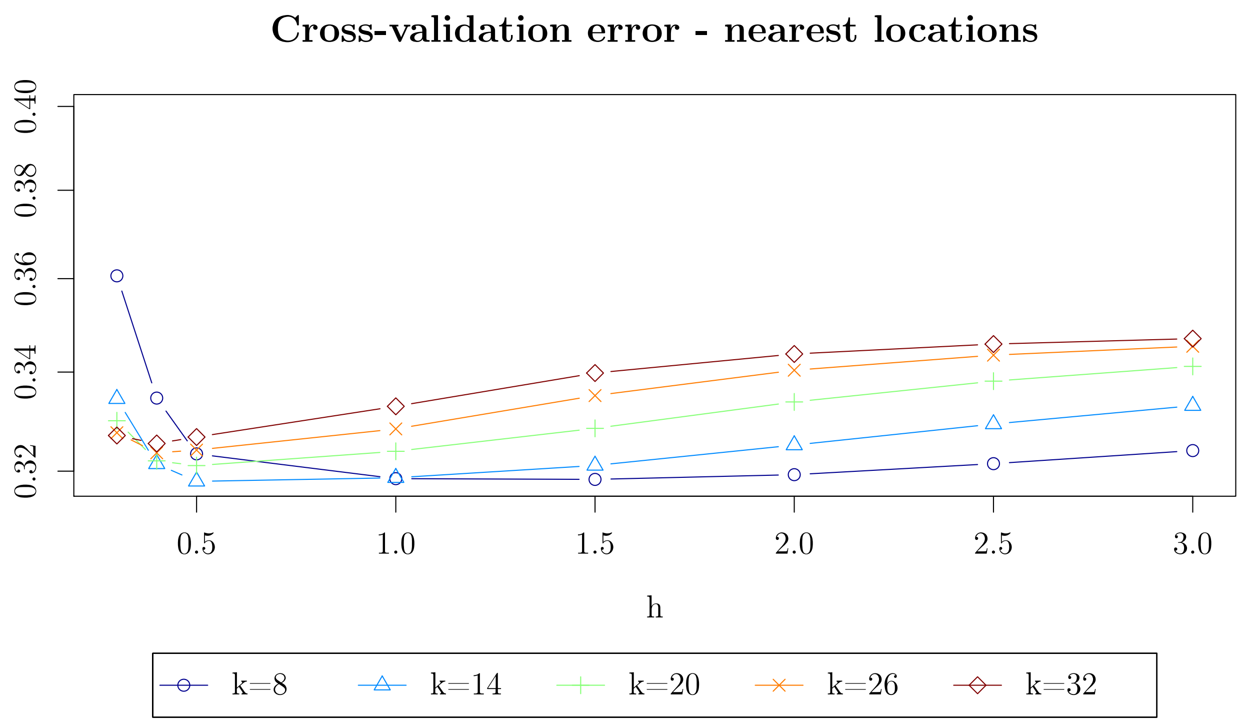
For the "class" dataset, the cross-validation errors different values of h and k can be found in Figure 3 (for the nearest locations bandwidth) and in Figure S1 of the Online Supplement (for the nearest observations bandwidth).

For the mean field, the nearest locations bandwidth yields the minimal cross-validation errors, with $h = 0.5$, $k = 14$. The nearest observations bandwidth yields a slightly higher minimal cross-validation error ($h = 1.5, k = 300$). As will be seen in Figures 4 and Figure S2 of the Online Supplement, the nearest location bandwidth yields maps that capture more of the variability of the mean MFCC field, whereas the nearest observations bandwidth seems to be oversmoothing. Therefore, we shall use the varying bandwidth with the nearest locations for the interpretation.

For the d_S-covariance field, the cross-validation curves decrease as the bandwidth parameter h increases, and seem to reach a plateau. We take the smallest value of h (and the corresponding k) that reaches the plateau, which is $h = 1, k = 32$ nearest locations. Although this choice seems to contradict Occam's razor (or "law of parsimony"), we make it consciously since we are not trying to prove the presence of spatial variation in the d_S-covariance field. It is an established fact in socio-linguistics that there is spatial variation in speech covariance (as can be seen in the vowel space analyses of Clopper et al. 2005, Strange et al. 2007, Clopper & Pierrehumbert 2008, Fox & Jacewicz 2017, Renwick & Olsen 2017). We are therefore trying to estimate it in the best possible way by choosing the most flexible model that fits the data, as long as it is as good as less variable models. While there are possible reasons why the spatial variation in d_S-covariance is not evident from the cross-validation curves, which could, for example, include the confounding effect of sex or age on the MFCCs, the microphone and room reverberation effect, or a small number of sound tokens where there is a mismatch between the geographical location of the recording and the spoken dialect of that region, we shall see in Section 4.3 that there is in fact evidence to support that the d_S-covariance field is not constant.

4.2 Projection of the mean field onto principal components

Visualization of the field of mean MFCC is not a straightforward task. Indeed, at each location x in Great Britain, $\hat{m}(x)$ is an element of $L^2\left([0,1],\mathbb{R}^p\right)$. A vizualization of the field $\hat{m}$ can be obtained by projection onto suitable elements of $L^2\left([0,1],\mathbb{R}^p\right)$, i.e. by looking at the map $x \mapsto \langle \hat{m}(x), \varphi \rangle$ for various $\varphi \in L^2\left([0,1],\mathbb{R}^p\right)$. Here we choose to project onto the principal components of the MFCCs $\{Y_{lj} : l = 1, \ldots, L; j = 1, \ldots, n_l\}$ (i.e. the pointwise multivariate PCA, that is, the multivariate PCA of $Y_{jl}(t)$ evaluated over a discrete grid of values t, or other words, our PCA is based on the eigen-analysis of the sample covariance matrix of $Y_{lj}(t)$—and not its correlation matrix—evaluated over a discrete grid of values t; another approach could be to use the method proposed in Chiou et al. (2014)). This allows the reproduction of the geographical variation of the projections which capture most of the variability in the original data and to compare the fields estimated for different values of h and k, the projection directions being independent from them.



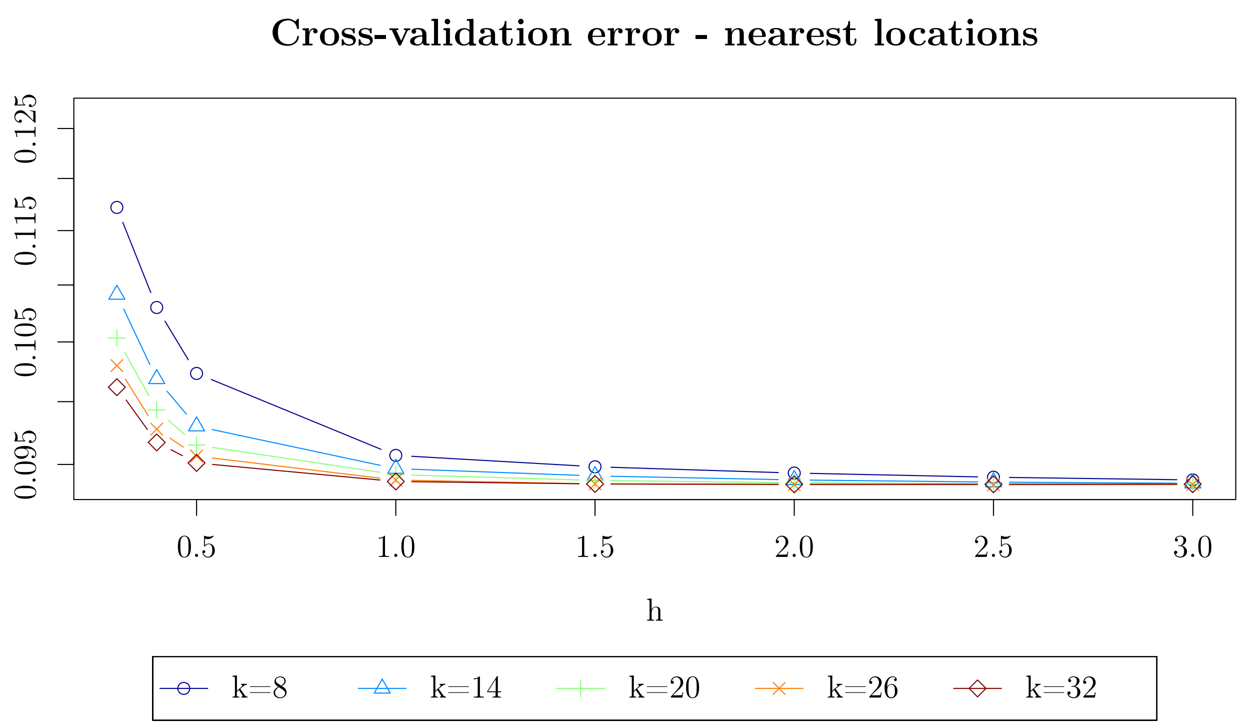


Figure 3: Cross-validation curves of the "class" dataset for the mean MFCC field (top) and the d_S-covariance field (bottom) when the bandwidth is adjusted using the k-th nearest locations.

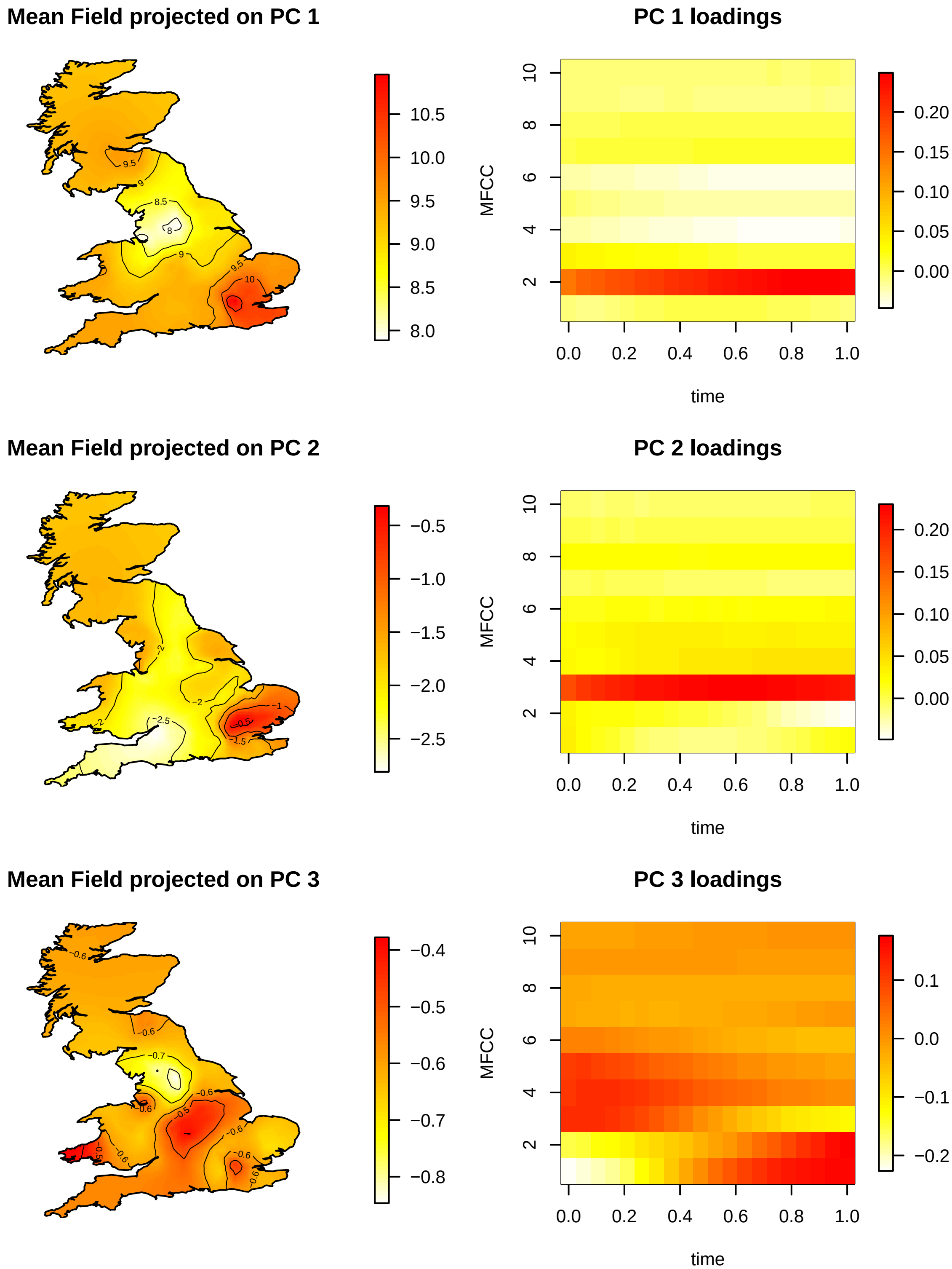


Figure 4: Left: Color maps with contours of the mean smooth MFCC field obtained for the "class" vowel with $h = 0.5$ and $k = 14$th nearest locations (denoted *NL map* in the text), projected onto the first three principal components directions (from top to bottom) of the original data $\{Y_{lj}(t) : l = 1, \dots, L; j = 1, \dots, n_l\}$. Right: Colour image representing the projection directions (loadings).

The maps of the projections of the estimated field for the choice of h and k that minimises the cross-validation error can be found Figure 4 (this corresponds to the nearest locations bandwidth, with $h = 0.5$, $k = 14$). The maps of projections for the nearest observations bandwidth can be found in Figure S2 of the Online Supplement. The first principal component direction (which accounts for 25.2% of the total variance) essentially considers the energy on the second cepstral coefficient (92.4% of its total energy), i.e. on the low frequencies. The second principal component (which accounts for 19.6% of the total variance) essentially considers the energy in the third cepstral coefficient (91.2% of its total energy), again energy in the low frequencies. The third principal component direction (which accounts for 8.8% of the total variance) mainly consists of time dynamics (along the sound length) of the relative volume, and the second and third cepstral coefficients, with some moderate time dynamics in the cepstral coefficients 4 and 5. The fact that most of the energy in the first and second PC loadings concentrate on a single cepstral coefficient (92.4% and 91.2% of their respective total energy) confirms that the MFCC representation is indeed a suitable representation for speech sounds. The map of the mean field projected into the first principal component direction highlights the difference between the region around London and the rest of the country, in particular part of North England (and most strongly around Bradford). The projection into the second principal component direction produces high values in East England, and contrasts these values with South West, West Midlands, Yorkshire and the Humber, and North East England, with the strongest contrast being with South West England. The projection in the third principal component direction produces low values in North England, and contrasts this region with isolated regions, such as East Midlands, the London area, and South West Wales. Figures S3 and S4 in the Online Supplement show the counties and regions of Great Britain, and are provided as geographical aids[1].

In order to assess whether there is spatial information in the mean field estimate, we compare our estimates with a simulation where the mean and the error terms have no spatial

[1]The maps shown in this paper do not include the Isle of Wight, located off the south coast of Great Britain, as there is no data present here, and as it is not simply connected to the rest of mainland UK, it is not possible to provide smooth estimates there.

information. The results of the simulation provide evidence in support of spatial structure for the mean field, which is expected since the cross-validation curves have a clear minimum. Details of the simulation are given in Section S5 of the Online Supplement.

4.3 MFCC d_S-Covariance Field

While the mean MFCC field captures the information about the average dialect sound changes, the regional variability of such dialect sounds may well also be of considerable interest. We therefore also want to explore how the d_S-covariance changes over the region of interest. While it is in principle possible to use dimension reduction methods, the interpretation of projections of the d_S-covariance may be problematic, as discussed in Section 3.1. An alternative way to represent the d_S-covariance variation is to consider a single location of interest and plot the square-root distances (averaged over the length of the sound) between the d_S-covariance at the location of interest, and the d_S-covariances at all other locations of the map. This produces 2D surfaces that reflect which parts of the country are more similar or dissimilar to the location of interest with respect to d_S-covariance. However, such maps are not directly interpretable, because many of their features appear due to the smoothing method. Indeed, Figure 5 shows the contours of the pairwise distances from Harlow (Essex), overlaid by contours obtained from 100 simulations from a model with constant mean and constant d_S-covariance field (details of the simulations are given in Section S5 of the Online Supplement), a procedure which can be considered a bootstrap approximation to the underlying null field. We can see in the Figure that the general form of the contours of the data and the simulations have similar shapes (such systematic effects are not present for the mean MFCC field, as can be seen from Figure S9 of the Online Supplement). This is because the raw d_S-covariances $\breve{\Omega}_l$ are quite noisy (indeed, Figure S8 in the Online Supplement, which shows the pairwise distance between the raw d_S-covariances against their corresponding geographical distance, has a nugget). Even though the contours of the data and the simulations have similar shapes, there are some significant differences between them. In Figure 5, we notice that as one moves away from Harlow (Essex), the distances between the d_S-covariances are growing faster in the data than what would be expected if there was no spatial structure in the d_S-covariance field (this can be seen by noticing that the thick dashed lines are not always contained in the bulk of the thin lines of the corresponding color). However, this is not true for all regions of Great Britain. Indeed, Figure 6, which shows the pairwise distances from Morecambe (Lancashire), does not exhibits such features as clearly. In principle, one could look at such maps of contours of distances from each region of Great Britain to assess whether or not the d_S-covariance field is varying spatially, but this is cumbersome and not visually appealing. A more appropriate tool for this purpose is to represent a normalized version $z_D(x, y)$ of the pairwise distances between d_S-covariances at locations x and y. The definition of $z_D(x, y)$ is as follows:

$$z_D(x, y) = \frac{D(x, y) - \overline{D^*}(x, y)}{\sigma^*(x, y)}, \tag{4.3}$$

where $D(x, y)$ is the distance between the d_S-covariances at x and y estimated from the data, $\overline{D^*}(x, y)$, respectively $\sigma^*(x, y)$, is the average, respectively the standard deviation, of $\{D^{*b}(x, y) : b = 1, \ldots, 100\}$, where $D^{*b}(x, y)$ is the distance between the d_S-covariances at x and y for the b-th simulation replicate (for both the data and all the simulations, the smoothing parameters are $h = 1$, $k = 32$ nearest locations). The notation $z_D(x, y)$ is chosen because (4.3) can be interpreted as a z-score for the distance between the d_S-covariances between locations x and y of the data, under the null hypothesis that the d_S-covariance field is constant. Figure 7 shows the surfaces $\{z_D(x_0, y) : y \in \mathcal{E}\}$ for the two locations $x_0 \in \mathcal{E}$ corresponding to the contours of Figures 5 and 6. We can see in Figure 7 that the values of $y \in \mathcal{E} \mapsto z_D(x_0, y)$, for x_0 corresponding to Harlow (Essex), are all larger than 2 for y in the Midlands and South of England, indicating a difference between their d_S-covariance and that of Harlow. For x_0 corresponding to Morecambe (Lancashire), the values of $y \in \mathcal{E} \mapsto z_D(x_0, y)$ are below 2 (and even negative) for y in North East England, East Midlands, and South East England, indicating little to no difference between their d_S-covariance and that of

Morecambe. These conclusions are coherent with (and make more precise) those made from Figures 5 and 6. Instead of looking at each surface $\{z_D(x_0, y) : y \in \mathcal{E}\}$ separately, it is possible to consider many such surfaces for locations all over Great Britain to get a global appreciation of the variation of the d_S-covariance field. Figure 8 shows the maps associated to many representative locations together with their geographical position in Great Britain, a "map of maps". We can see in the Figure that there is a very strong indication that the d_S-covariances of the region around Glasgow and Edinburgh are different from those of North England, and that the d_S-covariance of the Midlands are different from those in South and South-West England. There is also very strong indication that the d_S-covariance around Northamptonshire is different from those of East England and South-East England, and moderate to strong indication that the d_S-covariances of South England are different from those of the rest of England. All of these interpretations should be of course tempered by the fact that they are drawn from a very crude univariate representation of the d_S-covariance field (namely, the z-scores of their pairwise distances), and while it allows to find regions where the d_S-covariance field is varying spatially, it is not clear if a small value of the z-score $z_D(x, y)$ implies that there is no difference between the d_S-covariances at x and y. Figures S3 and S4 in the Online Supplement show the counties and regions of Great Britain, and are provided as geographical aids.

5 Discussion

We presented a method to explore spatial variation of sound processes which is of interest in particular for dialectology and comparative linguistics. The need to model the change in the covariability between frequencies, as well as in the mean sound, led us to propose the novel statistical concept of d-covariance, i.e. a definition of covariability that relies on a metric distance d different from the Euclidean (Frobenius) distance. This allows the use of metrics

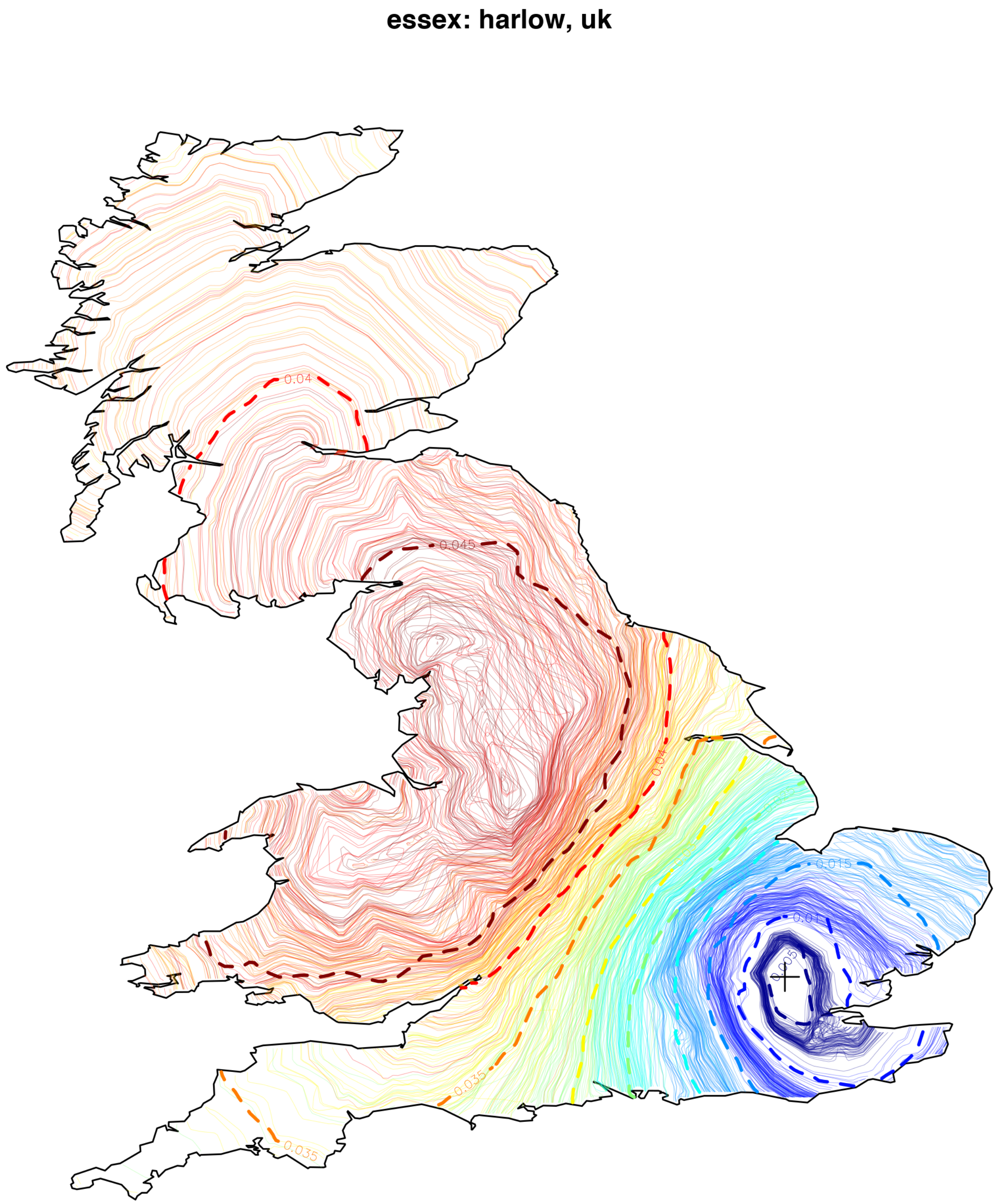


Figure 5: Contours of the pairwise distances between the d_S-covariances at Harlow (Essex), and other location in Great Britain. The thick dashed lines correspond to the contours (level sets) for the BNC data, and the corresponding contours for each of the 100 simulations are given in thin lines, with the corresponding color.

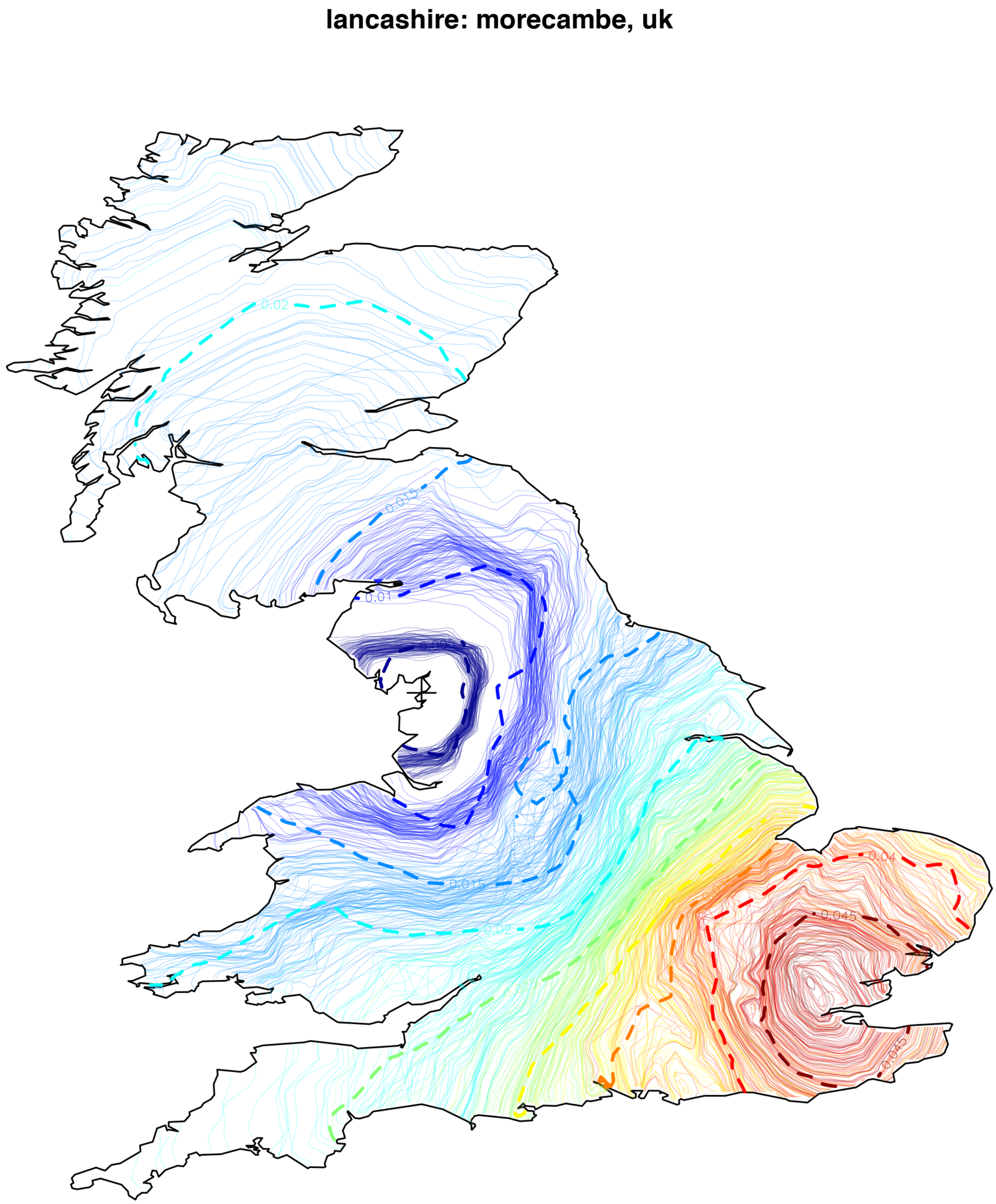


Figure 6: Contours of the pairwise distances between the d_S-covariances at Morecambe (Lancashire), and other location in Great Britain. The thick dashed lines correspond to the contours for the BNC data, and the corresponding contours for each of the 100 simulations are given in thin lines, with the corresponding color.

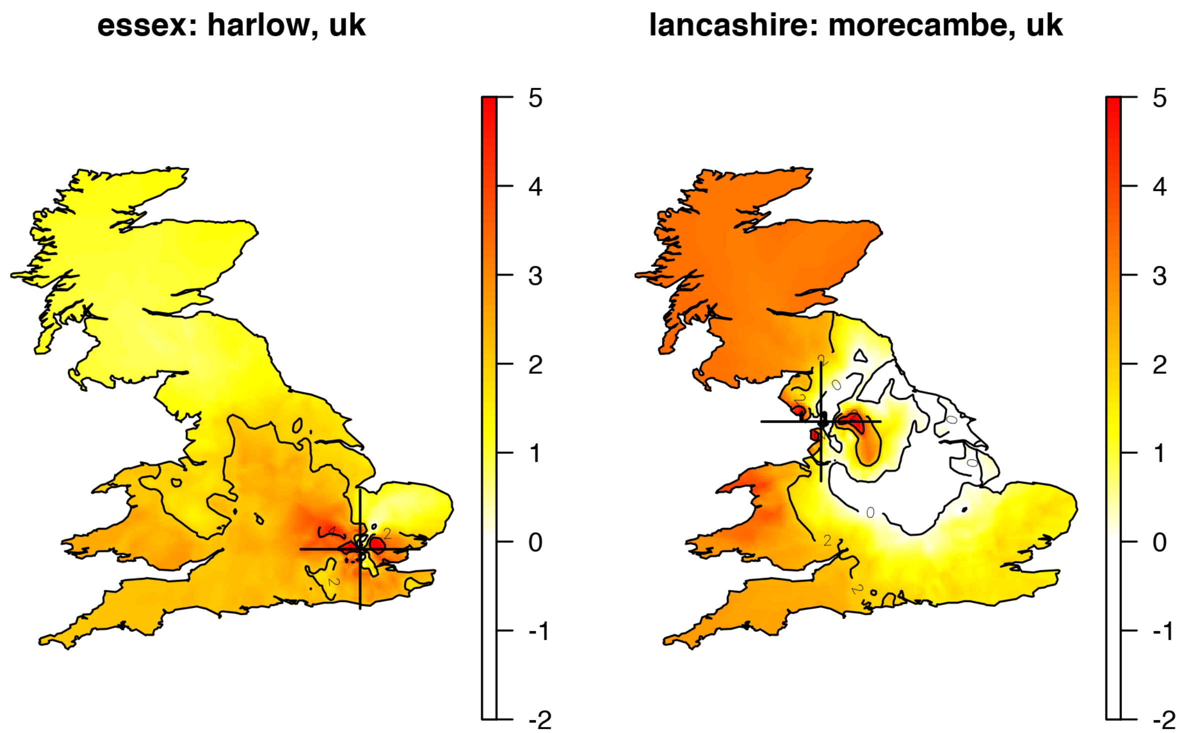


Figure 7: Z-scores $\{z_D(x_0, y) : y \in \mathcal{E}\}$ of the pairwise distances between the d_S-covariances from Harlow (Essex) (left), and from Morecambe (Lancashire) (right). The "+" on the maps represent the location of the corresponding x_0. Notice that the left sub-figure indicates a difference between the d_S-covariance of Harlow and those of the Midlands and South of England, whereas the right sub-figure indicates little to no difference between the d_S-covariance of Morecambe and those of North East England, East Midlands, and South East England. The definition of $z_D(x_0, y)$ is given in (4.3).

that do not produce swelling effects, while estimating the d-covariance consistently in the locations where observations are available. In particular, we chose the square root distance d_S described in Dryden et al. (2009) because it is defined for positive semi-definite matrices, and an explicit expression is available, as we showed in Section 3. It is clear that other metrics could be used within this framework, and indeed recent work on choosing metrics (Petersen & Müller 2016) and smoothing under general metrics (Petersen & Müller 2018) could prove relevant to this setting. However it is important to remember that the choice of metric should be considered within a data application context as well.

We used a Mel Frequency Cepstral Coefficients (MFCC) representation for the sound

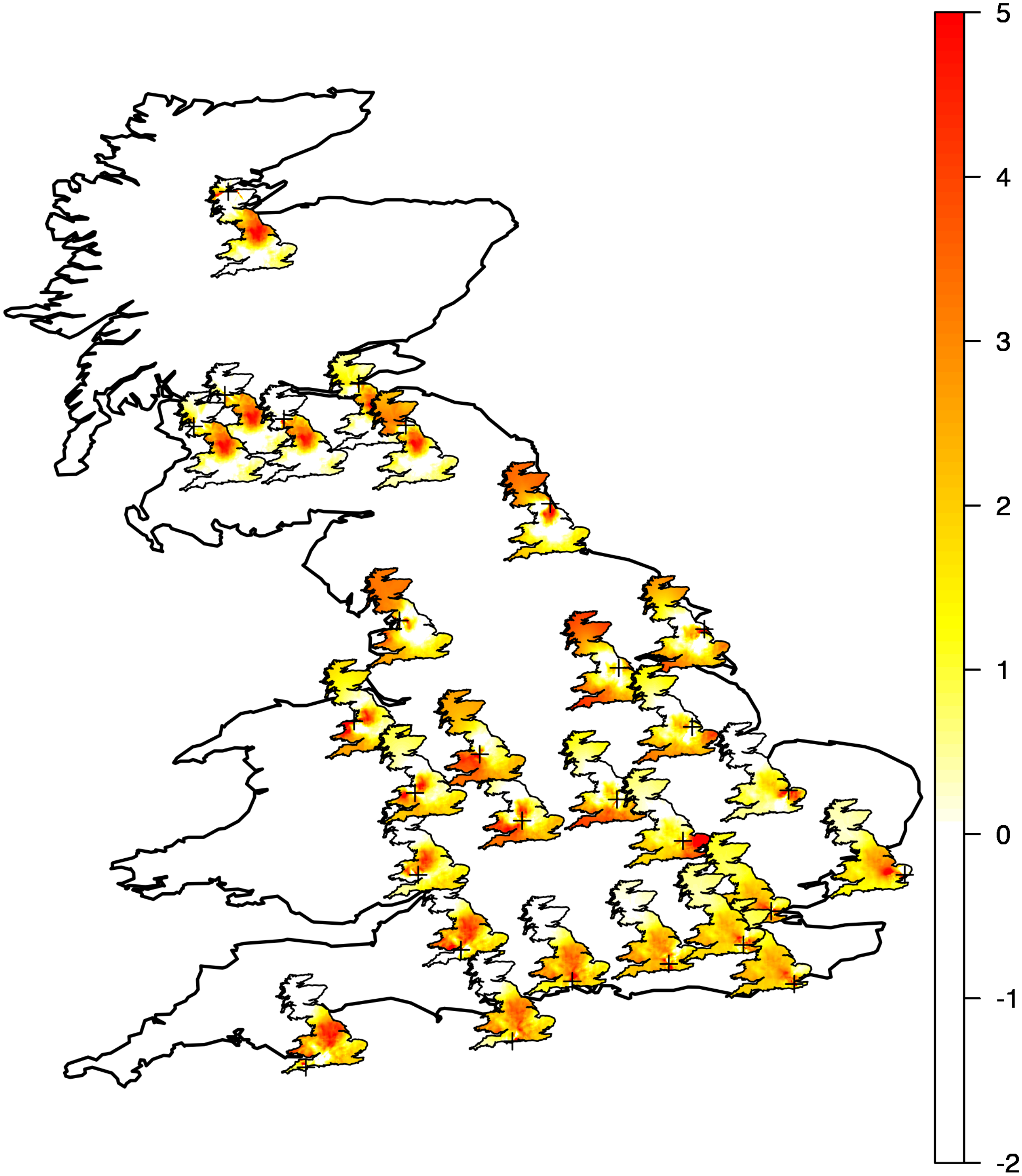


Figure 8: Z-scores $\{z_D(x_0, y) : y \in \mathcal{E}\}$ of the pairwise distances between the d_S-covariances for a number of representative locations x_0 in Great Britain, reported on the corresponding positions on the geographical man of Great Britain. The '+' on the small maps represent the location of the corresponding x_0. The definition of $z_D(x_0, y)$ is given in (4.3).

data objects because this has been found empirically to provide a better sound reconstruction, especially in the modified version of the algorithm proposed by Erro et al. (2014). Moreover, the fact that the frequency domain is partitioned into a relatively small number of channels (through a weighted averaging over a range of contiguous frequencies) makes this representation more robust to small frequency misalignments across speakers. MFCCs can be then treated as multivariate functional data, and we proposed a model where both the mean and the d_S-covariance between coefficients change smoothly in space. We proposed to estimate these smooth fields with a non-parametric estimator, and showed that this provides consistent estimates both for the mean and for the d_S-covariance field. We also integrated into the smoothing procedure a geographical distance based on the shortest path on the mesh used to triangulate the possibly non-convex region of interest. This required a non-trivial argument to show the consistency of the derived estimator and it has a wider applicability wherever there is the need of accounting for a complex geographical domain.

The proposed method allows, for the first time, the sound variation to be studied using speech recordings directly (as opposed to phonetic transcription), and provides a continuous model for the sound change (through its mean and d_S-covariance) in place of discrete regions boundaries, such as those traditionally reported in isoglosses. We analysed speech data from the spoken part of the British National Corpus, and focused on the pronunciation of the vowel in words such as *fast* or *class*, which is known to vary on a dialect basis (Upton & Widdowson 2013), and has particularly prominent variations in British English. While it is possible to listen to the reconstructed sounds (as given in the Supplementary Materials), visual maps are often useful to recognise both global patterns and local features. Exploring the estimated mean and d_S-covariance fields, we uncovered geographical patterns that resemble established findings about the vowel pronunciation (such as the contrast between the North and South-East England). However, the variation appears to be somewhat smoother than expected (i.e. from traditional dialectological maps of 'isoglosses'), to the point where it is possible to identify intermediate regions not easily classified by a hard clustering. This invites additional studies to explore other sounds and further exploration of this and alternative corpora.

Indeed, possible immediate extensions for this work include studying the joint behaviour of multiple words/sounds in the language and taking into account additional (non geographical) covariates, such as socio-economic variables.

SUPPLEMENTARY MATERIAL

Online Supplement: Details on the data preprocessing, simulation study, and example illustrating the advantage of d-covariances.

"class" data set and related functions: Data set and functions used in the illustration of our smoothing method in Section 4.

Sounds: Samples of the vowel sound data, effect of the PC loadings, and examples of reconstructed sounds.

References

Arsigny, V., Fillard, P., Pennec, X. & Ayache, N. (2006), 'Log-Euclidean metrics for fast and simple calculus on diffusion tensors', *Magnetic resonance in medicine* **56**(2), 411–421.

Arsigny, V., Fillard, P., Pennec, X. & Ayache, N. (2007), 'Geometric means in a novel vector space structure on symmetric positive-definite matrices', *SIAM Journal on Matrix Analysis and Applications* **29**(1), 328–347.

Aston, J. A., Chiou, J.-M. & Evans, J. P. (2010), 'Linguistic pitch analysis using functional principal component mixed effect models', *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **59**(2), 297–317.

Baíllo, A. & Grané, A. (2009), 'Local linear regression for functional predictor and scalar response', *Journal of Multivariate Analysis* **100**(1), 102–111.

Boj, E., Caballé, A., Delicado, P., Esteve, A. & Fortiana, J. (2016), 'Global and local distance-based generalized linear models', *Test* **25**(1), 170–195.

Boj, E., Delicado, P. & Fortiana, J. (2010), 'Distance-based local linear regression for functional predictors', *Computational Statistics & Data Analysis* **54**(2), 429–437.

Carmichael, O., Chen, J., Paul, D. & Peng, J. (2013), 'Diffusion tensor smoothing through weighted karcher means', *Electronic journal of statistics* **7**, 1913.

Chiou, J.-M., Chen, Y.-T. & Yang, Y.-F. (2014), 'Multivariate functional principal component analysis: A normalization approach', *Statistica Sinica* pp. 1571–1596.

Clopper, C. G. & Pierrehumbert, J. B. (2008), 'Effects of semantic predictability and regional dialect on vowel space reduction', *The Journal of the Acoustical Society of America* **124**(3), 1682–1688.
URL: *https://doi.org/10.1121/1.2953322*

Clopper, C. G., Pisoni, D. B. & de Jong, K. (2005), 'Acoustic characteristics of the vowel systems of six regional varieties of American English', *The Journal of the Acoustical Society of America* **118**(3), 1661–1676.
URL: *http://asa.scitation.org/doi/10.1121/1.2000774*

Coleman, J., Aston, J. & Pigoli, D. (2015), Reconstructing the sounds of words from the past, *in* The Scottish Consortium for ICPhS 2015, ed., 'Proceedings of the 18th International Congress of Phonetic Sciences', the University of Glasgow, Glasgow, UK. Paper number 0296.

Coleman, J., Baghai-Ravary, L., Pybus, J. & Grau, S. (2012), 'Audio bnc: the audio edition of the spoken british national corpus', *Phonetics Laboratory, University of Oxford* .
URL: *http://www.phon.ox.ac.uk/AudioBNC*

Cooley, J. W. & Tukey, J. W. (1965), 'An algorithm for the machine calculation of complex Fourier series.', *Math. Comput.* **19**, 297–301.

Delicado, P., Giraldo, R., Comas, C. & Mateu, J. (2010), 'Statistics for spatial functional data: some recent contributions', *Environmetrics* **21**(3-4), 224–239.

Dryden, I. L., Koloydenko, A. & Zhou, D. (2009), 'Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging', *The Annals of Applied Statistics* **3**(3), 1102–1123.

Erro, D., Sainz, I., Navas, E. & Hernáez, I. (2011), HNM-based MFCC+F0 extractor applied to statistical speech synthesis, *in* 'ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings', pp. 4728–4731.

Erro, D., Sainz, I., Navas, E. & Hernaez, I. (2014), 'Harmonics plus noise model based vocoder for statistical parametric speech synthesis', *IEEE Journal on Selected Topics in Signal Processing* **8**(2), 184–194.

Fan, J. & Gijbels, I. (1995), 'Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation', *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 371–394.

Fox, R. A. & Jacewicz, E. (2017), 'Reconceptualizing the vowel space in analyzing regional dialect variation and sound change in American English', *The Journal of the Acoustical Society of America* **142**(1), 444–459.

Francis, W. N. (1959), 'Some dialect isoglosses in england', *American Speech* **34**(4), 243–250.

Gervini, D. (2008), 'Robust functional estimation using the median and spherical principal components', *Biometrika* **95**(3), 587–600.

Gold, B., Morgan, N. & Ellis, D. (2011), *Speech and Audio Signal processing: Processing and Perception of Speech and Music*.
URL: *http://books.google.com/books?id=p0mNAF8gMmUC&pgis=1*

Gromenko, O., Kokoszka, P., Zhu, L. & Sojka, J. (2012), 'Estimation and testing for spatially indexed curves with application to ionospheric and magnetic field trends', *The Annals of Applied Statistics* **6**(2), 669–696.

Hadjipantelis, P. Z., Aston, J. A. & Evans, J. P. (2012), 'Characterizing fundamental frequency in mandarin: A functional principal component approach utilizing mixed effect models', *The Journal of the Acoustical Society of America* **131**(6), 4651–4664.

Hadjipantelis, P. Z., Aston, J. A., Müller, H.-G. & Evans, J. P. (2015), 'Unifying amplitude and phase analysis: A compositional data approach to functional multivariate mixed-effects modeling of mandarin chinese', *Journal of the American Statistical Association* **110**(510), 545–559.

Kraus, D. & Panaretos, V. M. (2012), 'Dispersion operators and resistant second-order functional data analysis', *Biometrika* **99**(4), 813–832.

Kretszchmar, W. A. (1996), 'Quantitative areal analysis of dialect features', *Language Variation and Change* **8**(01), 13–39.

Lehmann, C. (2004), 'Data in linguistics', *The Linguistic Review* **21**(3-4), 175–210.

Locantore, N., Marron, J. S., Simpson, D. G., Tripoli, N., Zhang, J. T. & Cohen, K. L. (1999), 'Robust principal component analysis for functional data (with discussions)', *Test* **8**(1), 1–73.
URL: *http://link.springer.com/10.1007/BF02595862*

Menafoglio, A. & Petris, G. (2016), 'Kriging for hilbert-space valued random fields: The operatorial point of view', *Journal of Multivariate Analysis* **146**, 84–94.

Nerbonne, J., Colen, R., Gooskens, C., Kleiweg, P. & Leinonen, T. (2011), 'Gabmap-a web application for dialectology', *Dialectologia: revista electrònica* pp. 65–89.

Nerbonne, J. & Kretzschmar, W. (2003), 'Introducing computational techniques in dialectometry', *Computers and the Humanities* **37**(3), 245–255.

Petersen, A. & Müller, H.-G. (2016), 'Fréchet integration and adaptive metric selection for interpretable covariances of multivariate functional data', *Biometrika* **103**(1), 103–120.

Petersen, A. & Müller, H.-G. (2018), 'Fréchet Regression for Random Objects with Euclidean Predictors', *The Annals of Statistics* **to appear**.

Pigoli, D., Aston, J. A., Dryden, I. L., Secchi, P. et al. (2014), 'Distances and inference for covariance operators', *Biometrika* **101**(2), 409–422.

R Core Team (2016), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Renwick, M. E. L. & Olsen, R. M. (2017), 'Analyzing dialect variation in historical speech corpora', *The Journal of the Acoustical Society of America* **142**(1), 406–421.

Sangalli, L. M., Ramsay, J. O. & Ramsay, T. O. (2013), 'Spatial spline regression models', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **75**(4), 681–703.

Srivastava, A., Wu, W., Kurtek, S., Klassen, E. & Marron, J. (2011), 'Registration of functional data using fisher-rao metric', *arXiv preprint arXiv:1103.3817* .

Strange, W., Weber, A., Levy, E. S., Shafiro, V., Hisagi, M. & Nishi, K. (2007), 'Acoustic variability within and across German, French, and American English vowels: Phonetic context effects', *The Journal of the Acoustical Society of America* **122**(2), 1111–1129.
URL: *https://doi.org/10.1121/1.2749716*

The Functional Phylogenies Group (2012), 'Phylogenetic inference for function-valued traits: speech sound evolution', *Trends in Ecology & Evolution* **27**(3), 160–166.

Tucker, J. D., Wu, W. & Srivastava, A. (2013), 'Generative models for functional data using phase and amplitude separation', *Computational Statistics & Data Analysis* **61**, 50–66.
URL: *http://www.sciencedirect.com/science/article/pii/S0167947312004227*

Upton, C. & Widdowson, J. D. A. (2013), *An Atlas of English Dialects: Region and Dialect*, Routledge.

Wang, H., Marron, J. et al. (2007), 'Object oriented data analysis: Sets of trees', *The Annals of Statistics* **35**(5), 1849–1873.

Wood, S. N., Bravington, M. V. & Hedley, S. L. (2008), 'Soap film smoothing', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**(5), 931–955.

Wu, W. & Srivastava, A. (2014), 'Analysis of spike train data: Alignment and comparisons using the extended fisher-rao metric', *Electronic Journal of Statistics* (1), 1776–1785.

Yuan, Y., Zhu, H., Lin, W. & Marron, J. S. (2012), 'Local polynomial regression for symmetric positive definite matrices', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **74**(4), 697–719.